Remigiusz Marciniak 244781
Zadanie 3: Dopasowanie lokalne par sekwencji

Repozytoria

https://gitlab.com/RemigiuszMMarciniak/pythonlocalsequencealignmentproject

1.
Złożoność obliczeniowa czasowa i pamięciowa zaimplementowanych algorytmów
Złożoność czasowa

Złożoność czasowa algorytmu, który generuje tablicę punktów lokalnego dopasowania wynosi $O(n^2)$, ponieważ algorytm ten jest uzależniony od rozmiarów macierzy n oraz m. Natomiast złożoność algorytmu, który zwraca optymalną ścieżke jest zależna od długości tej ścieżki, więc jest stałą O(1).

Złożoność pamięciowa

Złożoność pamięciowa algorytmu, który generuje lokalne dopasowanie wynosi $O(n^2)$, ponieważ dane są zapisywane w macierzy n x m. Natomiast złożoność pamięciowa algorytmu, który zwraca optymalną ścieżkę wynosi O(n), ponieważ w pamięci ścieżka zapisywana jest w dwóch tablicach jednowymiarowych o długości n.

2.
Instrukcja użytkownika

Aby uruchomić program należy posiadać zainstalowanego Pythona w wersji 3 na swoim komputerze. Aby uruchomić program należy wprowadzić następującą komendę:

python3 main.py

Następnie po spacji należy dodać argumenty. Pierwszym z nich jest typ wczytywania sekwencji – występuje manual, file oraz server. W przypadku użycia manual przykładowa komenda wygląda następująco:

python3 main.py manual ATATAGC seq1 GGTATAC seq2 -2 test1 test1 score_matrix.txt

Po manual wprowadzamy sekwencje pierwszą, a następnie określamy jej etykiete, analogicznie postępujemy z 2gą sekwencją. Następnie określamy punktacje za przerwe, nadajemy nazwę plików oraz na samym końcu wczytujemy macierz punktacji. Macierz punktacji musi być określona w sposób ściśle określony, w przeciwnym wypadku program się nie uruchomi.

W przypadku file komenda wygląda następująco:

python3 main.py file file1.txt file2.txt seq2 -2 test2 test2 score_matrix.txt

Natomiast przy użyciu server komenda wygląda następująco:

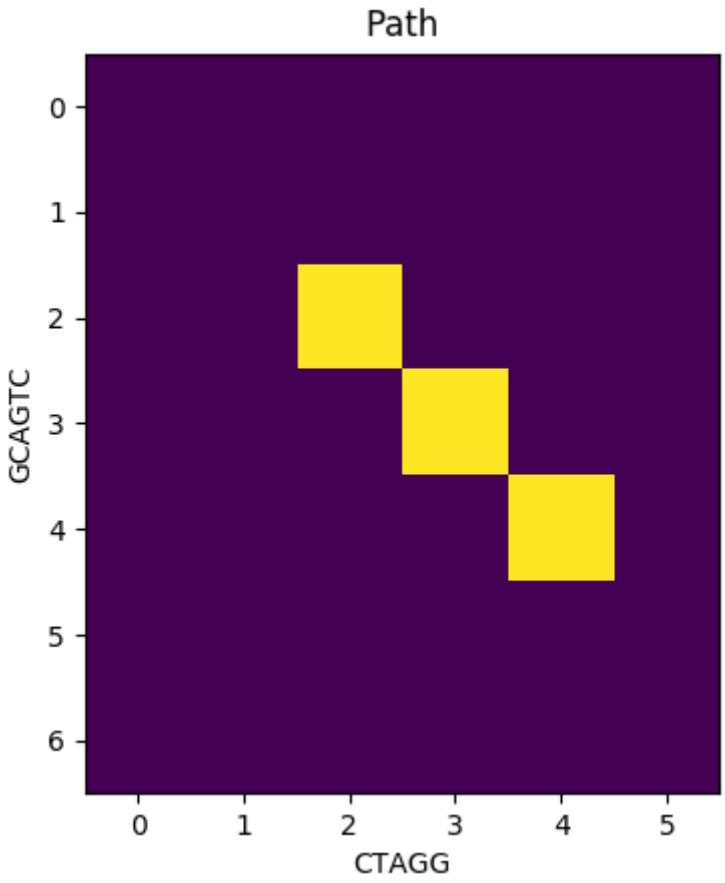
python3 main.py server url1 url2 -2 test3 test3 score_matrix.txt

W przypadku gdy używamy Linuxa możemy przy użyciu ./tests.sh uruchomić zestaw testów, który wygeneruje kilka przykładowych dopasowań. Ponadto, możemy dodać swoje testy przy użyciu basha.

3.
Porównanie przykładowych par sekwencji ewolucyjnie powiązanych i niepowiązanych
Przykładowe dopasowania

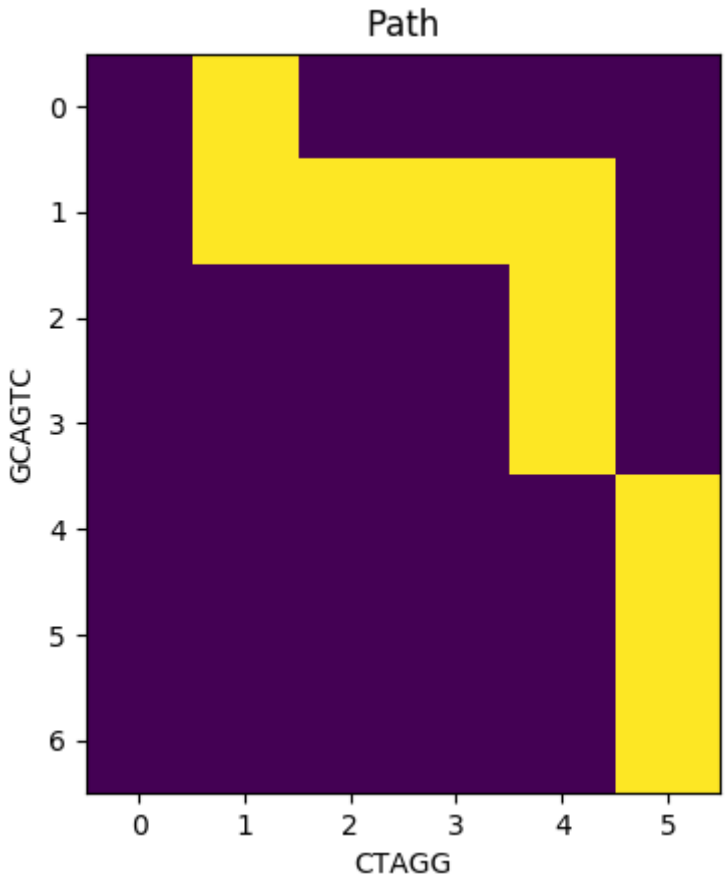
```
#manual
python3 main.py manual GCAGTC 1st CTAGG 2nd -2 t1 t1 SM.txt
python3 main.py manual GCAGTC 1st CTAGG 2nd -2 t2 t2 SMMATCH2MISMATCH0.txt
python3 main.py manual GCAGTC 1st CTAGG 2nd 1 t3 t3 SMMATCH2MISMATCH0.txt
python3 main.py manual GCAGTC 1st CTAGG 2nd -1 t4 t4 SMMATCH2MISMATCH5.txt
python3 main.py manual GCAGTC 1st CTAGG 2nd -1 t5 t5 SMMATCH1MISMATCH-1.txt
python3 main.py manual GCAGTC 1st CTAGG 2nd -1 t6 t6 SM.txt
```

Poniżej zostaną przedstawione wyniki dla t1,t3,t4.



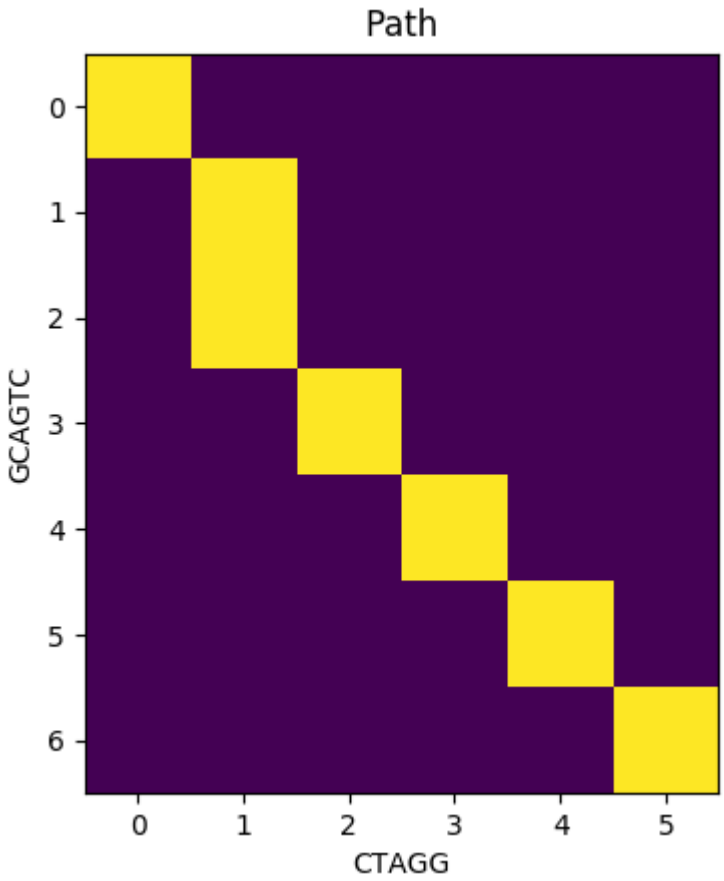

```
>seq1 1st
GCAGTC
>seq2 2nd
CTAGG
>optimal path
AG
||
AG
```

```
# - C T A G G
- 0 0 0 0 0 0
G 0 0 0 0 2 2
C 0 2 0* 0 0 0
A 0 0 0 2* 0 0
G 0 0 0 0 4* 2
T 0 0 2 0 2 0
C 0 2 0 0 0 0
```




```
>seq1 1st
GCAGTC
>seq2 2nd
CTAGG
>optimal path
G---CAGTC
xxxxxx|xx
-TAG--G--
```

```
# - C T A G G
- 0 0* 0 0 0 0
G 0 1* 2* 3* 4* 5
C 0 2 3 4 5* 6
A 0 3 4 5 6* 7
G 0 4 5 6 7 8*
T 0 5 6 7 8 9*
C 0 6 7 8 9 10*
```



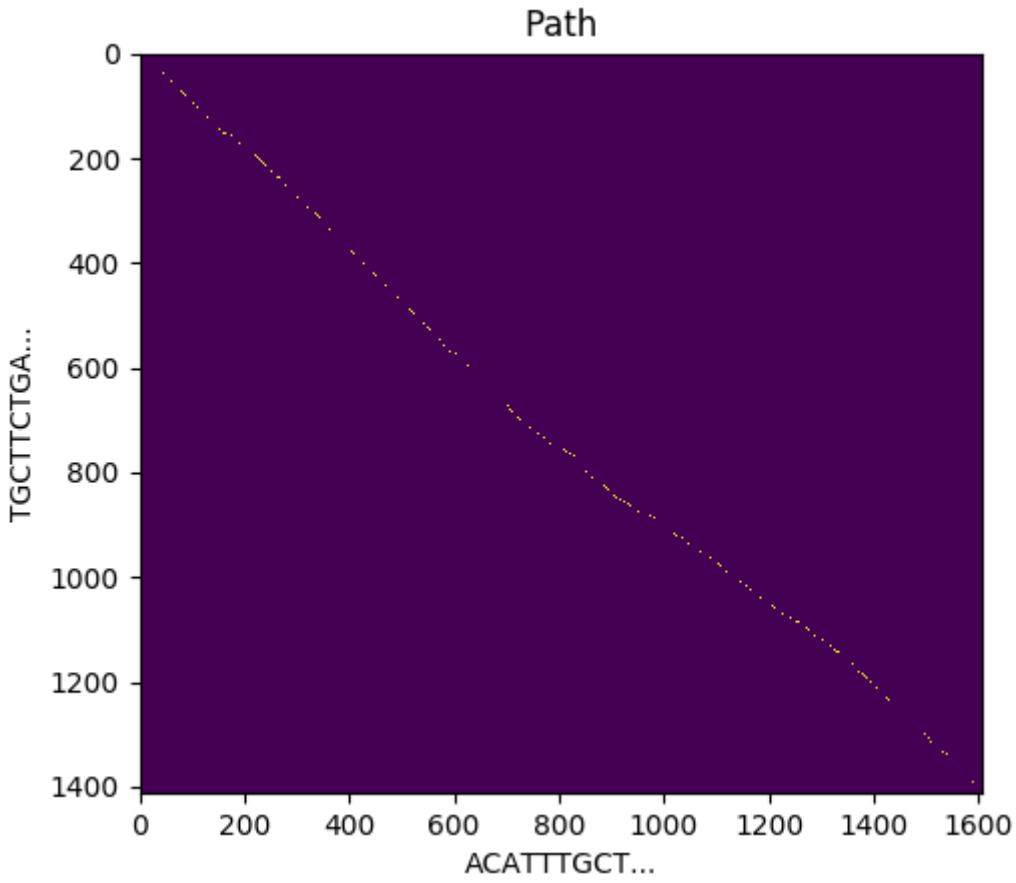

```
>seq1 1st
GCAGTC
>seq2 2nd
CTAGG
>optimal path
GCAGTC
xxxxxx
C-TAGG
```

#	-	C	T	A	G	G
-	0*	0	0	0	0	0
G	0	5*	5	5	4	3
C	0	4*	10	10	10	9
A	0	5	9*	12	15	15
G	0	5	10	14*	14	17
T	0	5	9	15	19*	19
C	0	4	10	14	20	24*

```
#real data
python3 main.py file RNHGBB1.txt HSHGBB1.txt -2 t7 t7 SM.txt
python3 main.py file RNHGBB1.txt HSHGBG1.txt -2 t8 t8 SM.txt
python3 main.py file HSHGBG1.txt HSHGBB1.txt -2 t9 t9 SM.txt
```

Porównanie par sekwencji ewolucyjnie powiązanych

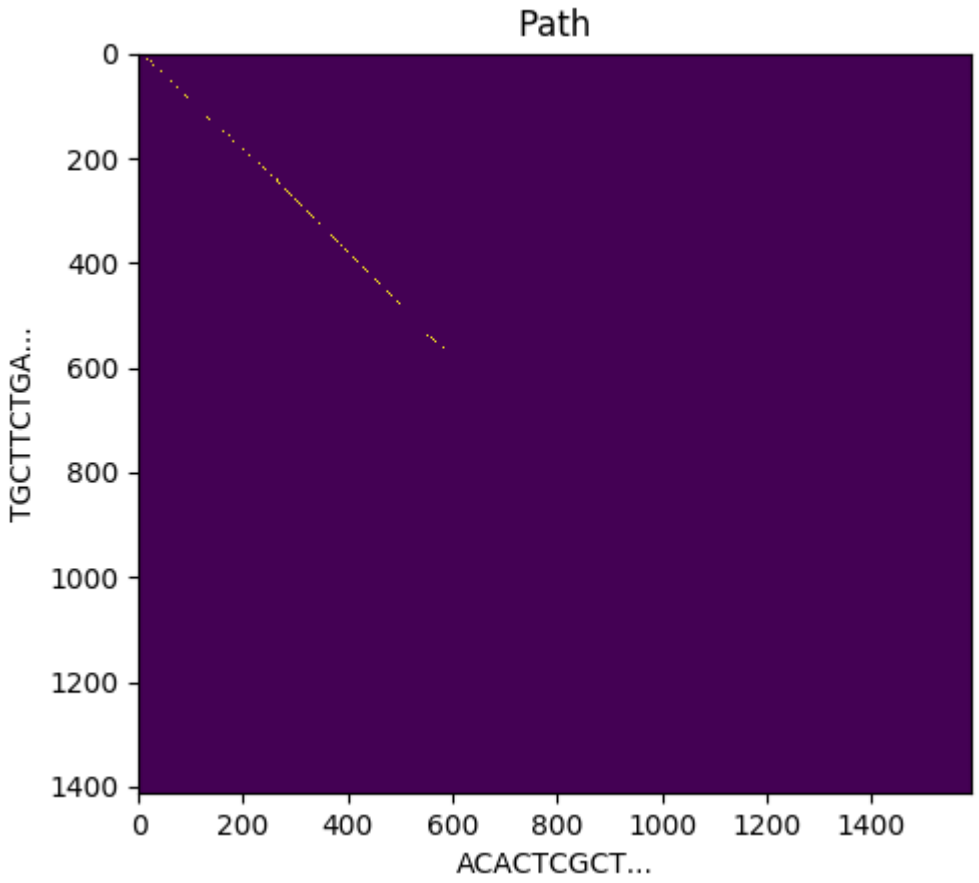
Porównane zostały geny kodujące podjednostkę beta hemoglobiny człowieka i szczura.




```
>seq1 >NC_005100.4:c168972680-168971269 Rattus norvegicus strain mixed chromosome 1, Rnor_6.0
['T', 'G', 'C', 'T', 'T', 'C', 'T', 'G', 'A', 'C', 'A', 'T', 'A', 'G', 'T', 'T', 'G', 'T', 'G', 'T', 'T', 'G', 'A', 'C',
>seq2 >NC_000011.10:c5227071-5225464 Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly
['A', 'C', 'A', 'T', 'T', 'T', 'G', 'C', 'T', 'T', 'C', 'T', 'G', 'A', 'C', 'A', 'C', 'A', 'A', 'C', 'T', 'G', 'T', 'G',
>optimal path
TGCTTCTG--ACATA-GTTGTGTT-GACTCA-CAA-ACTCAGAAACAGACACCATGGTGCA-CCTGA---CTGATGCTGAGAAG-GCTG-CTGTTA-ATG-GCCTGTGGGG-AAAGGTGAA
||||||||xx|||x|xxx|||||xx|||x|x|||xx|||xx||||||||||||||||||xx||||xxx||||x|xx||||||xx|||x|x||||xx||xx|||||||||xx||||||||
TGCTTCTGACACA-AC--TGTGTTC-ACT-AGCAAC-CTC--AAACAGACACCATGGTGCAT-CTGACTCCTGA-G--GAGAAGT-CTGCC-GTTAC-TGC-CCTGTGGGGC-AAGGTGAA
```

Porównanie par sekwencji ewolucyjnie niepowiązanych

Porównany został gen kodujący podjednostkę beta hemoglobiny szczura oraz gen kodujący podjednostkę gamma hemoglobiny człowieka.



```
>seq1 >NC_005100.4:c168972680-168971269 Rattus norvegicus strain mixed chromosome 1, Rnor_6.0
['T', 'G', 'C', 'T', 'T', 'C', 'T', 'G', 'A', 'C', 'A', 'T', 'A', 'G', 'T', 'T', 'G', 'T', 'G', 'T', 'T', 'G', 'A', 'C',
>seq2 >NC_000011.10:c5249857-5248269 Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly
['A', 'C', 'A', 'C', 'T', 'C', 'G', 'C', 'T', 'T', 'C', 'T', 'G', 'G', 'A', 'A', 'C', 'G', 'T', 'C', 'T', 'G', 'A', 'G',
>optimal path
GCTTCT-G-ACATAGT-T--GTGTTGACTCACA-AA-CT-C-AG--AAACAGAC-ACCAT-GGTGCA---C-C-TGA-CTGATGCTGA-GAAGGCT-GCT-GT---TAATGGCCTGTGGGG
||||||x|x||xxx||x|xx|x|||x|x|||x|x||x|x||xxxxx|||||xx||||x|||x||xxx|x|xx||xxx||x|xx||xx||||||xx||xx|xxxx||xx|||||||||
GCTTCTGGAAC---GTCTGAG-GTT-A-TCA-ATAAGCTCCTAGTC---CAGACG-CCATGGGT-CATTTCACA-GAG--GA-G--GAC-AAGGCTA-CTA-TCAC-AA--GCCTGTGGGG
```

```
# - A C A C T C G C T T C T G G A A C G T C T G A G G T T A T C A A T A A G C T C C T A
- 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
T 0 0 0 0 0 0 2 0* 0 0 2 2 0 2 0 0 0 0 0 0 2 0 2 0 0 0 0 2 2 0 2 0 0 0 2 0 0 0 0 2 0 0 2 0
G 0 0 0 0 0 0 0 0 2* 0 0 0 0 0 4 2 0 0 0 2 0 0 0 4 2 2 2 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0
C 0 0 2 0 2 0 2 0 4* 2 0 2 0 2 0 0 0 2 0 0 2 0 2 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 4 2 2 2 0 0
T 0 0 0 0 0 0 4 2 0 2 6* 4 2 4 2 0 0 0 0 0 0 2 0 4 2 0 0 0 2 2 0 2 0 0 0 2 0 0 0 2 6 4 2 4 2
T 0 0 0 0 0 0 2 0 0 0 4 8* 6 4 2 0 0 0 0 0 0 2 0 2 0 0 0 0 2 4 2 2 0 0 0 2 0 0 0 0 4 2 0 4 2
C 0 0 2 0 2 0 4 2 2 2 2 6 10* 8 6 4 2 0 2 0 0 4 2 0 0 0 0 0 2 0 0 4 2 0 0 0 0 0 2 2 6 4 2
T 0 0 0 0 0 0 4 2 0 0 4 4 8 12* 10* 8 6 4 2 0 2 2 6 4 2 0 0 2 2 0 2 2 0 0 2 0 0 0 0 4 4 2
G 0 0 0 0 0 0 2 0 4 2 2 2 2 6 10 14 12* 10* 8 6 4 2 0 4 8 6 4 2 0 0 0 0 0 0 0 0 0 2 0 2 2
A 0 2 0 2 0 0 0 2 0 0 0 4 8 12 10 14 12* 10 8 6 4 2 6 10 8 6 4 2 2 0 0 2 2 0 2 2 0 0 0 0
C 0 0 4 2 4 2 2 0 4 2 0 2 6 10 8 12 10 14* 12 10 8 6 4 8 6 4 2 0 0 0 2 0 0 0 0 0 0 2 0 2
A 0 2 2 6 4 2 0 0 2 0 0 0 4 8 6 10 14 12* 10 8 6 4 2 6 4 2 0 0 2 0 0 4 2 0 2 2 0 0 0 0 0
T 0 0 0 4 2 6 4 2 0 4 2 0 2 6 4 8 12 10* 8 12 10 8 6 4 2 0 4 2 0 4 2 2 0 4 2 0 0 0 2 0 0
A 0 2 0 2 0 4 2 0 0 2 0 0 0 4 2 6 10 8* 6 10 8 6 4 8 6 4 2 0 4 2 0 4 4 2 6 4 2 0 0 0 0 0
G 0 0 0 0 0 0 2 0 4 2 0 0 0 0 2 6 4 8 6 10* 8 6 4 8 6 10 8 6 4 2 0 0 2 2 0 4 2 6 4 2 0 0 0
T 0 0 0 0 0 0 2 0 2 0 4 2 0 2 0 4 2 6 4 8 12* 10* 8 6 4 8 6 10 8 6 4 2 0 0 4 2 0 4 2 6 4 2
T 0 0 0 0 0 0 2 0 0 0 2 6 4 2 0 2 0 4 2 6 10 8 12* 10* 8* 6 4 8 12 10 8 6 4 2 2 0 0 2 0 4
G 0 0 0 0 0 0 0 0 2 0 0 4 2 0 4 2 0 2 0 4 8 6 10 14 12 10* 8 6 10 8 6 4 2 0 0 0 0 2 0 2 0
T 0 0 0 0 0 0 2 0 0 0 2 2 0 4 2 0 0 0 0 2 6 4 8 12 10 8* 6 10 8 6 10 8 6 4 2 0 0 0 0 2 0 0
G 0 0 0 0 0 0 0 0 2 0 0 0 0 2 6 4 2 0 0 2 4 2 6 10 8 12 10* 8 6 4 8 6 4 2 0 0 0 2 0 0 0 0
T 0 0 0 0 0 0 2 0 0 0 2 2 0 2 4 2 0 0 0 0 4 2 4 8 6 10 8 12* 10 8 6 4 2 0 4 2 0 0 0 2 0 0
T 0 0 0 0 0 0 2 0 0 0 2 4 2 2 2 0 0 0 0 0 2 0 4 6 4 8 6 10 14* 12 10 8 6 4 2 0 0 0 0 2 0 0
G 0 0 0 0 0 0 0 0 2 0 0 2 0 0 4 4 2 0 0 2 0 0 2 6 4 6 10 8 12* 10 8 6 4 2 0 0 0 2 0 0 0 0
A 0 2 0 2 0 0 0 0 0 0 0 0 0 2 2 6 4 2 0 0 0 0 4 8 6 8 6 10 14* 12 10 8 6 4 2 2 0 0 0 0 0
C 0 0 4 2 4 2 2 0 2 0 0 2 0 0 0 4 2 6 4 2 2 0 2 6 4 6 4 8 12* 10 14 12 10 8 6 4 2 2 0 2
T 0 0 2 0 2 6 4 2 0 4 2 0 4 2 0 2 0 4 2 6 4 4 2 4 2 4 8 6 10 14* 12 10 8 12 10 8 6 4 4 2
```

Jak możemy zauważyć w pierwszym przypadku została wygenerowana optymalna sekwencja o długości większej niż w przypadku drugim, gdzie optymalna sekwencja zajmowała tylko początkową część obu sekwencji. Dzięki tej obserwacji być może jesteśmy w stanie stwierdzić, że w pierwszym przypadku obie sekwencje są ze sobą ewolucyjnie powiązane, zaś w drugim są niepowiązane(lub powiązane w małym stopniu).